Assessing the predictive performance of probabilistic caries risk assessment models: the importance of calibration.

Trottini M[1], Campus G[2,3*], Corridore D[4], Cocco F[3], Cagetti MG[5], Vigo MI[6], Polimeni A[4], Bossú M[4]

[1]Department of Mathematics, University of Alicante, Spain
[2]Clinic for Restorative, Preventive and Pediatric Dentistry, University of Bern, Switzerland
[3]Department of Surgery, Microsurgery and Medicine Sciences, School of Dentistry, University of Sassari, Italy
[4]Department of Oral and Maxillofacial Sciences, 'Sapienza' University of Rome, Rome, Italy
[5]Department of Biomedical, Surgical and Dental Science, University of Milan, Italy
[6]Department of Applied Mathematics, University of Alicante, Spain

Short title: Assessment of P-CRA models: the importance of calibration.
Keywords: caries risk assessment, Cariogram, calibration, discrimination

* Corresponding author
Prof. Guglielmo Campus
Clinic for Restorative, Preventive and Pediatric Dentistry
Zahnmedizinische Kliniken Bern (zmk bern)
University of Bern, Switzerland
Freiburgstrasse 7, CH-3010 Bern (Switzerland)
Tel (office) +41 (0)31 6322571
Tel (cellular) +41-(0)798441703
E-mail: Guglielmo.campus@zml.unibe.ch

Abstract

Probabilistic caries risk assessment models (P-CRA), such as the Cariogram, are promising tools to planning treatments in order to control and prevent caries. The usefulness of these models in informing patients and in medical decision making depends on two properties of the models known as discrimination and calibration. Current assessment of P-CRA models, however, ignores calibration and, as such, could be very misleading. The aim of this paper is provide tools for a proper assessment of the P-CRA models calibration and for its improvement when lacking. A combination of standard calibration tools (calibration plot, calibration in-the-large and calibration slope) with three novel measures of calibration (the Calibration Index and two related metrics E50 and E90) is proposed to establish if a model is well calibrated. Lack of calibration is addressed using Platt scaling and Isotonic regression. The proposed approach is described using data from a published follow-up study performed on children evaluated through a reduced Cariogram model. For this data a lack of calibration is detected. The Cariogram overestimates the actual risk of new caries for forecast probabilities <0.5 and underestimates the risk for forecast probabilities >0.6. Both Platt scaling and Isotonic regression significantly improve the calibration of the reduced Cariogram model, preserving its discrimination properties. The average specificity and sensitivity for both Platt scaling and Isotonic regression using as cut-off point 0.5 were greater than 83 and their sum well exceeds 160. The benefits of the proposed calibration methods are promising, but still further research in this field should be done.

Introduction

Current assessment and reporting of the performance of probabilistic caries risk assessment (P-CRA) models as predictor of new caries, relies on measures such as, sensitivity, specificity, accuracy (or misclassification error), positive and negative predictive values [Tellez et al., 2013; Senneby et al., 2015; Trottini et al., 2015; Cagetti et al., 2018]. However, these measures address only one dimension of the model's predictive performance, the discrimination, but ignore another important dimension, the calibration. Discrimination refers to the ability of the model to separate the instances (subjects) in which a certain event (development of new caries in the near future) will occur from those instances in which the event will not occur. Calibration, instead, refers to the statistical consistency between the forecast probabilities and the observed relative frequencies [Schmid and Griffit, 2005]. A predictive model is well calibrated if among the subjects for which the model returns a probability p that the event will take place, approximately p% of the subjects do in fact experience the event. Good calibration of a P-CRA model would imply that among those subjects for which the model forecasts a probability of new caries of 0.10, approximately 10% of the subjects develop new caries and so on. On the other hand, the lack of calibration of a P-CRA model in a certain study (even when the model has perfect discrimination) would indicate that the forecast probabilities of new caries do not match the observed frequencies of new caries. In this case the forecast probability does not correctly quantify the actual chances of new caries in the near future (for example subjects with a very low forecast probability of new caries lesions might have very high chances of experiencing new caries). As noted in [Van Calster et al. 2016] "For informing patients and medical decision making, calibration is the primary requirement" [Steyerberg et al., 2010], in addition "If the model is not […] well

calibrated, it must be regarded as not having been validated"[…]. To evaluate the classification performance […] is inappropriate" [Kim and Simon, 2011].

As indicated in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [Collins et al. 2015, Moons et al. 2015], calibration and discrimination should be reported in all prediction papers [see explanation of item 10d on page w33 in Moons et al. 2015]. However, this is not the case for P-CRA models.

A list of prospective studies that have been conducted to assess the performance of the P-CRA models as predictor of the binary event, "new caries in the near feature", is shown in Table 1. The table was build using the most recent review in [Cagetti et al., 2018] and adding two follow-up studies [Hayes et al., 2017; Garg et al., 2018]. Only the papers describing the results at follow-up have been included in the table, while prospective studies that do not focused on caries increment have been excluded. All the papers include some measure of discrimination. None of them, however, provides calibration measures or relevant data for a rigorous assessment of calibration. Thus, according to the existing guidelines for clinical prediction models, current assessment of the performance of P-CRA model is severely incomplete since ignores calibration and, as such, could be very misleading.

The aim of this paper is to provide tools and guidelines for: (i) a proper assessment of the P-CRA models calibration; and (ii) improve P-CRA models calibration if lack of calibration is detected.

A combination of standard calibration tools described in existing TRIPOD guidelines for clinical prediction models with a set of novel tools introduced in [Austin and Steyerberg,

2019] is proposed to assess calibration. The relevance of complementing standard measures of calibration with the novel tools is also discussed.

To deal with the lack of calibration of P-CRA models it is proposed to use the Platt scaling [Platt, 1999] and the Isotonic regression [Zadrozny and Elkan, 2002] to adjust the model's forecast probabilities improving the match between forecast probabilities and the observed frequencies of new caries. The proposed approach is illustrated using published data [Campus et al., 2009; 2012].

Material and Methods

Material

Both calibration assessment and recalibration methods are illustrated using the data from a previous follow-up study. The study population consists of 957 school children, aged 7-9 years, and resident in Sassari, the largest district in northern Sardinia (Italy), examined in 2009 [Campus et al., 2009]. A reduced Cariogram model without two saliva parameters was used to evaluate forecast probabilities of new caries in the near future. After two years 862 children (90%) from the original sample were re-examined and the actual caries increment for the permanent dentition was calculated [Campus et al., 2012].

Assessment of calibration

As a visual assessment of calibration the smoothed version of the calibration plot is considered [Austin and Steyerberg, 2014a]; the plot is produced by charting the forecast probabilities of new caries *versus* the smoothed observed probabilities (the *calibration curve*) together with the corresponding pointwise confidence intervals and the ideal 45º line of perfect calibration. Five measures are used to characterize the calibration curve

and to quantify its deviation from the ideal 45º line: calibration in-the-large (α), calibration slope (β) [Moons et al. 2015], and the Calibration index (ICI) with the companion metrics E50 and E90 [Austin and Steyerberg, 2019]. Perfect calibration corresponds to α=0, and β=1 and to ICI=E50=E90=0 (for the interpretation of these five measures please refers to Appendix A in the supplementary material).

Recalibration Methods

The Platt scaling [Platt, 1999] and the Isotonic regression [Zadrozny and Elkan, 2002] were used to address poor calibration. In the Platt scaling the function that transforms the forecast probabilities in order to match the observed frequencies is a parametric sigmoid function. The Isotonic regression [Zadrozny and Elkan, 2002] is a form of non-parametric regression, where the function used to transform the forecast probabilities is a non-decreasing step function. The measures of discrimination will be preserved in both transformations. Platt scaling and Isotonic were preferred to standard calibration methods (intercept and slope updating) because more suitable for the Cariogram model (see discussion).

Statistical analysis

The calibration plots have been obtained by regressing the binary outcome (presence/absence of new caries) on the forecast probabilities of new caries using the LOESS (Locally Estimated Scatterplot Smoothing) method. Corresponding pointwise confidence intervals have been obtained using bootstrap [Austin and Steyerberg 2014b]. Calibration slope, β, has been estimated by fitting a logistic model where the binary outcome is the response variable and the log odds of the forecast probabilities is the only predictor. The calibration in-the-large, α, has been estimated fitting the same logistic

model under the constraint that $\beta=1$, taking α as the model estimated intercept and the corresponding Wald's confidence intervals have been calculated. The ICI, E50 and E90 have been evaluated as the mean, the median and the 90^{th} percentile of absolute differences between the observed and the forecast probabilities of new caries. The span parameters for the LOESS was set to 0.75 [Austin and Steyberg, 2014b]. The Isotonic regression based on pair-adjacent violators algorithm (PAV) was used. The parameters of the sigmoid functions in the Platt scaling were estimated from the data using standard results from logistic regression [Platt, 1999]. To prevent overfitting Platt scaling/Isotonic regression functions were estimated in 1000 bootstrap samples and applied to the original data producing, for each recalibration method, 1000 sets of recalibrated probabilities. In order to appreciate the effect of recalibration methods in terms of discrimination, standard quantitative measures of discrimination such as the area under the ROC curve (AUC) were calculated; sensitivity (Se), specificity (Sp) and the accuracy or misclassification error (A) [Schmid and Griffit, 2005] were also calculated for the original data and the recalibrated data. The Brier score has been reported as overall measure of the predictive performance of the model.

Data were analysed using the free software environment for statistical computing and graphics R, version 3.6.1 [R core Team, 2013]. In particular, Optimal Youden cut-off, sensitivity, specificity and area under the ROC curve have been obtained using the Optimal Cutpoints package [López et al., 2014]; for (smoothed) calibration plots, calibration-in-the-large, calibration-slope, the val.prob.ci.2 which is an adaptation of the val.prob function in the rms package was used [Van Calster et al. 2016]; the ICI, E50 and E90 metrics were evaluated using the R code in [Austin and Steyerberg, 2019]; for the Isotonic regression and the Platt scaling, the CORElearn package [Robnik-Silkonja and

Savicky, 2019] and the standard logistic regression function from the stats package were used.

Results

The LOESS calibration plot of the data is shown in Figure 1a with the confidence band represented as grey shaded area. The plot identifies two regions in the range of the forecast probabilities with a lack of Cariogram's calibration, the region (0 / 0.5) and the region (0.6 / 0.8). The value of the calibration slope is β=1.973 and the corresponding lower bound of the 95% confidence interval is 1.723, indicating that for the forecast probabilities, low predictions are too high and high predictions are too low. As it is evident from the plot, the Cariogram overestimates the actual risk of new caries for a forecast probabilities <0.5, and underestimates the actual risk of new caries for a forecast probabilities >0.6. The calibration in-the-large is α=-0.318 with the upper bound of the 95% confidence interval for α is -0.152, indicating that the average forecast probability of new caries (0.41) is greater than the observed incidence of new caries at follow-up (0.36). The value of the ICI index, ICI=0.106, can be interpreted as the average weighted-distance from the ideal 45º line of the points in the calibration curve (Figure 1a). The E90 metric is 0.191 and the E50 index is 0.089.

The predictive performance of the original Cariogram model and its calibrated versions (Platt scaling and Isotonic regression) are compared in terms of standard measures of discrimination, calibration and overall performance in Table 2. The values reported in the table are average of the 1000 bootstrap estimates with the corresponding standard deviation in parenthesis. Very similar results (not shown) are obtained using three-fold cross validation. Both calibration methods were able to preserve quite well the good

discrimination properties of the Cariogram model. The average specificity (Se^{05}), sensitivity (Se^{05}), and misclassification error ($A^{0.5}$) for both the Platt scaling and the Isotonic regression (using as cut-off point 0.5) as well as the area under the ROC curve (AUC) are very similar to the corresponding values for the original Cariogram model .

On average, the ICI, E50 and E90 decrease by a factor of at least 4 when the original forecast probabilities are re-calibrated using the Platt scaling and the Isotonic regression. The Brier score in both cases reduces from 0.106 to 0.095. There is no evidence that one of the recalibration method outperforms the other. The difference between the calibration metrics for the Platt scaling and the Isotonic regression are within the error bars, *i.e.* within $\pm$ two standard deviations. As an alternative and complementary way to visualize the results in Table 2, Figures 1b and Figure 1c show the LOESS smoothed calibration plot for the calibrated probability obtained applying the Platt scaling and the Isotonic regression. In both cases, the calibration plot refers to the calibrated probabilities from a "representative" bootstrap sample: from the 1000 bootstrap samples, those producing calibration metrics close to the average values reported in Table 2 were selected. In both cases the 95% confidence intervals for the calibration in-the-large and the calibration slope contain the ideal values (0 and 1, respectively). In addition, the calibration curve is very close to the ideal 45º line and the confidence bands for the calibration curve contains the ideal 45º line across the whole range of predicted probabilities.

Discussion

The usefulness of P-CRA models as a tool to predict new caries, critically depends on a dimension of the predictive performance known as calibration. Current assessment of

these models, however, does not address calibration and as such, could be very misleading. The aim of the present paper was to overcome this limitation presenting a set of tools for: (1) a proper assessment of the P-CRA models calibration; and (2) for improvement of calibration when lack of calibration is detected.

This paper is based on data from a previous follow-up study on Sardinian school children with a reduced Cariogram model [Campus et al. 2009, 2012].

Despite the good discrimination property of the Cariogram in this study, lack of calibration was detected. The Cariogram overestimates the actual risk of new caries when it outputs forecast probabilities <0.5, and underestimates the actual risk of new caries when it outputs forecast probabilities >0.6. Both the Platt scaling and the Isotonic regression significantly improved the calibration. It is important to underline that calibration was not mentioned in the original paper [Campus et al. 2012] and this limitation is common to all published research on predictive performance of P-CRA models (for more details see appendix B in the supplementary material).

Four aspects of the proposed methodology deserve some further discussion :

- the choice of the calibration metrics;
- the version of the calibration plot used;
- the choice of the recalibration methods and the interpretation of the results that these methods produce;
- the relevance of the proposed methodology in terms of internal and external model's validation.

The motivation for reporting ICI, E50 and E90, in addition to the standard calibration measures (the calibration in-the-large, α, and calibration slope, β), is that these metrics

are able to complete the information provided by the standard measures (calibration in-the-large and slope). When the calibration curve deviates from the ideal 45º line, ICI, E50 and E90 would provide a numerical summary that allows to detect this deviation and the comparison of calibration across models [Austin & Steyerberg, 2019].

As an alternative to the LOESS calibration plot used in this paper, a standard calibration plot could be used. This plot is obtained by stratifying the sample on the basis of the forecast probability (subjects with similar forecast probabilities of new caries belong to the same *stratum*) and plotting the mean predicted probabilities *per stratum versus* the observed frequency of new caries *per stratum*. Although common, the grouping used in the standard calibration plot is somewhat arbitrary and imprecise [Steyerberg et al. 2010]. The smoothed calibration plot overcomes these limitations as it does not require to choose the number of groups. Assessment of calibration, in addition, is straightforward [Austin and Steyerberg 2014a,b].

In the LOOES, the presence of regions where the bootstrap based confidence intervals do not contain the 45º line highlights the lack of calibration. The distance of the confidence band from the 45º lines indicates how much the model is not calibrated.

Lack of calibration is a common problem. The calibration performance of a model developed using data from a certain population/geographical area, usually decreases when the model is applied to another population/geographical area. When this is the case, recalibration methods can be used to adjust the predictive model in order to improve the match with the observed probabilities. Standard recalibration methods include more or less extensive adjustment of the models' parameters [Moons et al. 2015]. Such an approach, however, is not feasible for P-CRA tools that do not fit a parametric model to the data such as the Cariogram. This is why in the paper the more generally applicable

methods Platt scaling and Isotonic regression, that directly transform's the model's forecast probabilities rather than the model's parameters, were preferred.

The results from the Platt scaling and the Isotonic regression need to be considered with attention. As described in this paper, the lack of calibration means that, for the Sardinian sample, the Cariogram's forecast probabilities should be adjusted. Both methods adjust the P-CRA model's probabilities to match the observed frequencies, so quantifying the actual chances of new caries; as a consequence, an overfitting could occur. Bootstrap was used to address this issue. Ideally, an independent study should be used to quantify the benefits of the two recalibration methods.

Within clinical prediction model, a clear distinction is necessary between internal validation of the model, as part of model development, and external validation, as the evaluation of the performance of the model in other data different from those used to develop the model. Calibration is a key aspect of both [Moons et al. 2015]. The methodology proposed in this paper to assess calibration can be used both for the internal and external validation of a predictive model. For the internal validation, however, techniques such as bootstrap and cross validation should be used [Moons et al. 2015]. Clearly, the proposed recalibration methods only make sense as part of external validation.

The present work has important research and clinical implications.

From a research perspective, it is able to identify and solve an important weakness of current assessment of P-CRA models: the absence of calibration. This paper should raise the proper level of awareness among researchers with the proposed methodology as the standard for assessment of P-CRA models calibration in caries research.

From a clinical perspective, the proposed assessment of calibration is crucial for informing patients on her/his real risk of development disease and planning suitable preventive and treatment plan.

Current understanding of both calibration properties of the P-CRA models and potential benefits of recalibration methods is still very limited and further research in this field should be performed.

References

Austin PC, Steyerberg EW: Bootstrap confidence intervals for loess-based calibration curves. Stat Med 2014b; 33:699-700.

Austin PC, Steyerberg EW: Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med 2014a;33:17-35.

Austin PC, Steyerberg EW: The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Medi 2019; 38:51-65.

Cagetti MG, Bontà G, Cocco F, Lingstrom P, Strohmenger L, Campus G: Are standardized caries risk assessment models effective in assessing actual caries status and future caries increment? A systematic review. BMC Oral Health 2018; 18:123.

Campus G, Cagetti MG, Sacco G, Benedetti G, Strohmenger L, Lingström P: Caries risk profiles in Sardinian schoolchildren using Cariogram. Acta Odontol Scand 2009; 67: 46–52.

Campus G, Cagetti MG, Sale S, Carta G, Lingström P: Cariogram validity in schoolchildren: a two-year follow-up study. Caries Res 2012;46:16–22.

Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55-63.

Dou L, Luo J, Fu X, Tang Y, Gao J, Yang D: The validity of caries risk assessment in young adults with past caries experience using a screening Cariogram model without saliva tests. Int Dent J 2018;68:221-26.

Gao XL, Di Wu I, Lo EC, Chu CH, Hsu CY, Wong MC. Validity of caries risk assessment programmes in preschool children. J Dent 2013;41:787–95.

Gao XL, Hsu C-YS, Xu Y, Hwarng HB, Loh T, Koh D. Building caries risk assessment models for children. J Dent Res 2010;89:637–43.

Garg A, Madan M, Dua P, Saini S, Mangla R, Singhal P, et al: Validating the usage of Cariogram in 5- and 12-year-old school-going children in Paonta Sahib, Himachal Pradesh, India: a 12-month prospective study. Int J Clin Pediatr Dent 2018;11:110-15.

Hayes M, Da Mata C, McKenna G, Burke FM, Allen PF: Evaluation of the Cariogram for root caries prediction. J Dent 2017;62:25-30.

Kim KI, Simon R: Probabilistic classifiers with high-dimensional data. Biostatistics 2011; 12:399-412.

Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW et al.: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:1-73.

Petersson GH, Fure S, Bratthall D: Evaluation of a computer-based caries risk assessment program in an elderly group of individuals. Acta Odontol Scand 2003;61:164–71.

Petersson GH, Isberg PE, Twetman S: Caries risk assessment in school children using a reduced Cariogram model without saliva tests. BMC Oral Health 2010b;10:5.

Petersson GH, Isberg PE, Twetman S: Caries risk profiles in schoolchildren over 2 years assessed by Cariogram. Int J Paediatric Dent 2010a;20:341–46.

Petersson GH, Twetman S: Caries risk assessment in young adults: a 3-year validation of the Cariogram model. BMC Oral Health 2015;15:17.

Platt J: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, in: Smola A, Bartlett P, Schölkopf B, Schuurmans D, editors. Advances in Large Margin Classifiers, Cambridge MA: MIT Press; 1999; p. 61–74.

R Core Team R. A language and environment for statistical computing, in R Foundation for Statistical Computing, Vienna, Austria, 2013. [checked 11/20/2019] Available from: http://www.R-project.org.

Schmid CH, Griffith JL: Multivariate Classification Rules: Calibration and Discrimination, in: Armitage P, Colton T, editors. Encyclopedia of Biostatistics, Chichester: Wiley; 2005. p. 3491–97.

Senneby A, Mejare I, Sahlin NE, Svensäter G, Rohlin M§: Diagnostic accuracy of different caries risk assessment methods. A systematic review. J Dent 2015;43:1385–93.

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et.: Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21:128-138.

Sudhir KM, Karthik KK, Fareed N, Srikanth E, Nelagondanahalli T: Validation of Cariogram as a tool for caries risk prediction among 12-year- old institutionalized children - a longitudinal follow-up study. J Oral Health Dent Manag 2017;16:1-8. • Austin PC, Steyerberg EW: Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med 2014a; 33:517-535..

Tellez M, Gomez J, Pretty I, Ellwood R, Ismail AI: Evidence on existing caries risk assessment systems: are they predictive of future caries? Community Dent Oral Epidemiol 2013;41:67–78.

Trottini M, Bossù M, Corridore D, Ierardo G, Luzzi V, Saccucci M, Polimeni A: Assessing risk factors for dental caries: a statistical modeling approach. Caries Res 2015;49:226–35.

Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina M, Steyerberg EW: A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016; 74:167-176.

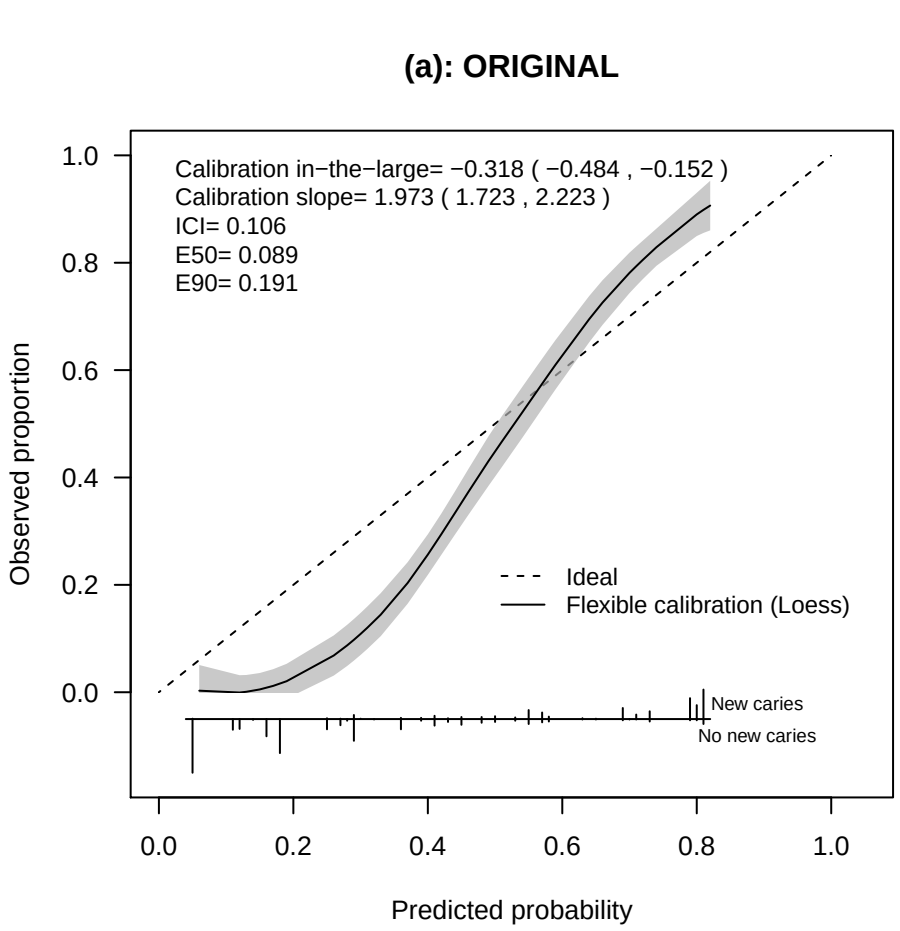
Zadrozny B, Elkan C: Transforming classifier scores into accurate multiclass probability estimates, in: Hand D, Keim DA, Raymond NG, editors. Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2002; p. 694–699.

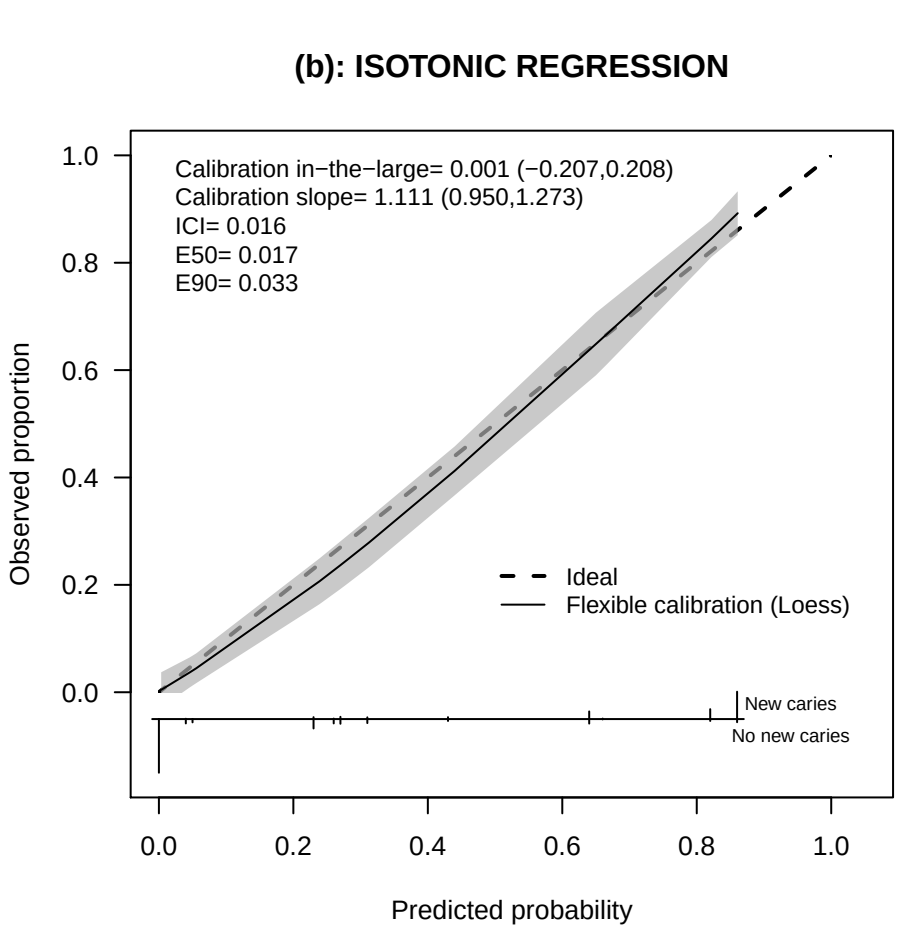
Legends:

Figure 1: Calibration plot for the original forecast probabilities [Campus et al. 2012] (a), and the recalibrated probabilities using Isotonic regression (b) and Platt scaling (c). The LOESS smoothed calibration curve is represented as a solid line in black; the ideal 45º line as a dashed line, and the 95% confidence band as a shaded grey area. The relative frequency distribution of the forecast probabilities of new caries for children that at follow-up did and did not developed new caries are represented as rescaled histograms at the bottom of each plot (labeled as "New caries", and "No new caries"). For Platt scaling and Isotonic regression, the calibration plot refers to the calibrated probabilities from one "representative" bootstrap sample.

Table 1. Prospective studies that have been performed to assess the performance of P-CRA models as predictor of the binary event "New caries in the near feature". C=Comprehensive; S=Screening. Measures of discrimination reported: area under the ROC curve (AUC), sensitivity (Se), specificity (Sp), Youden index (Se +Sp), accuracy (A), positive predictive values (PPV), negative predictive values (NPV). NUS-CRA denotes the National University of Singapore Caries Risk Assessment model.

Table 2. Discrimination measures (dark grey area), calibration measures (light grey area) and Brier score (BS) for the original Cariogram model and its calibrated versions using Platt scaling and Isotonic regression. AUC denotes the area under the ROC curve. For discrimination, $Se^{0.5}$, $Sp^{0.5}$ and $A^{0.5}$ denote the sensitivity, specificity and accuracy associated with the standard cut-off 0.5. ICI is the Calibration Index. For Platt Scaling and Isotonic regression reported values are average of the 1000 bootstrap estimates with the corresponding standard deviation in parenthesis.


(a): ORIGINAL
Calibration in−the−large= −0.318 (−0.484 , −0.152)
Calibration slope= 1.973 (1.723 , 2.223)
ICI= 0.106
E50= 0.089
E90= 0.191
Ideal
Flexible calibration (Loess)
New caries
No new caries
Observed proportion
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0


(b): ISOTONIC REGRESSION
Calibration in−the−large= 0.001 (−0.207,0.208)
Calibration slope= 1.111 (0.950,1.273)
ICI= 0.016
E50= 0.017
E90= 0.033
Ideal
Flexible calibration (Loess)
New caries
No new caries
Observed proportion
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0

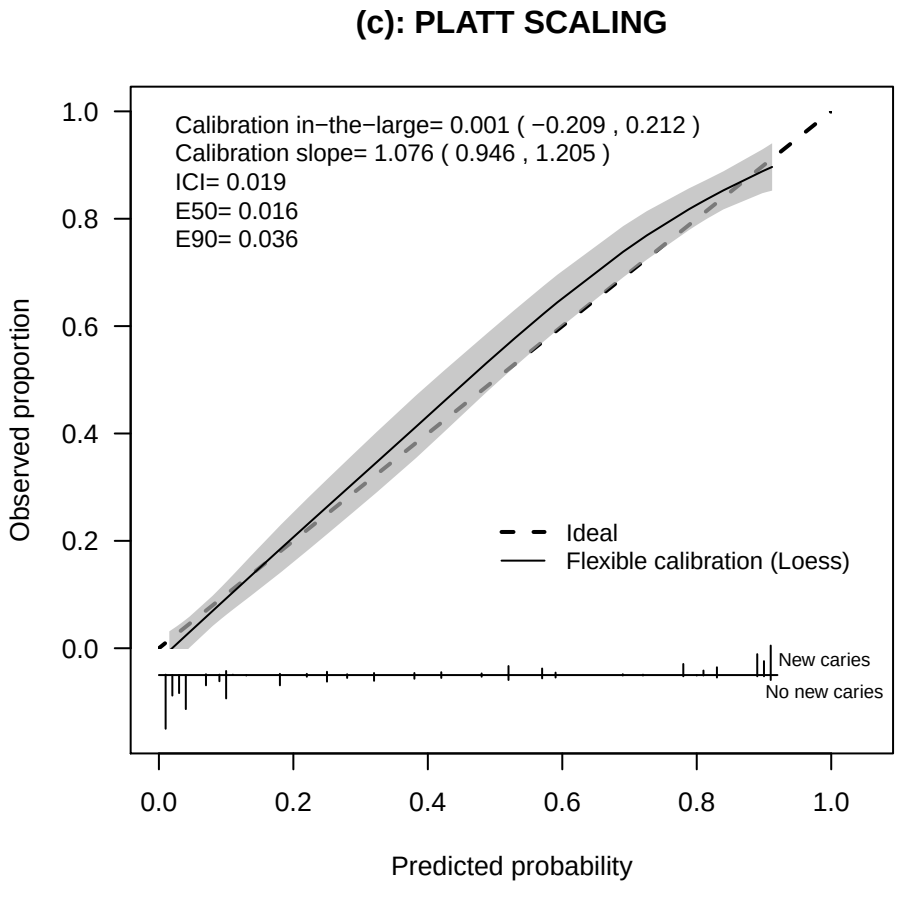

(c): PLATT SCALING
Calibration in−the−large= 0.001 (−0.209 , 0.212)
Calibration slope= 1.076 (0.946 , 1.205)
ICI= 0.019
E50= 0.016
E90= 0.036
Ideal
Flexible calibration (Loess)
New caries
No new caries
Observed proportion
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0

Table 1.

	Authors (year)	P-CRA model (version).	Sample size (% dropout)	Age	Study time (years)	Definition of new caries	Measures of discrimination reported	Cariogram categories	Observed relative frequency (f_j) by cariogram category	Sample size (N_j) by cariogram category
Preschool children	Gao et al. (2013)	Cariogram, NUS-CRA (S,C)	485 (10.8%)	3	1	∆dmft>0	AUC, Se, Sp, Se+Sp, A	5	**YES**	**YES**
	Gao et al. (2010)	Cariogram, NUS-CRA (S,C)	1576 (11.6%)	3-6	1	∆dmft>0	AUC, Se, Sp, Se+Sp, A	Not used	NO	NO
	Garg et al. (2018)	Cariogram (C)	250 (3.8%)	5	1	∆dmft>0	None	3	**YES**	**YES**
School children	Campus et al. (2012)	Cariogram (S)	861 (10%)	7-9	2	∆DFS>0	AUC, Se, Sp,	5	NO	**YES**
	Petersson et al.(2010a)*	Cariogram (C)	392 (10.5%)	10-11	2	∆DMFS>0	AUC, Se, Sp, PPV, NPV**	5	**YES**	**YES**
	Petersson et al. (2010b)	Cariogram (S)	392 (10.5%)	10-11	2	∆DMFS>0	AUC, Se, Sp, PPV, NPV	5	**YES**	**YES**
	Garg et al. (2018)	Cariogram (C)	249 (4.2%)	12	1	∆DMFT>0	None	3	**YES**	**YES**
	Sudhir et al. (2017)	Cariogram (C)	36 (0%)	12	1.5	∆DMFT>0	AUC, Se, Sp, A, PPV, NPV,	5	**YES**	**YES**
Young Adults	Petersson et al. (2015)	Cariogram (C)	982 (24.2%)	19	3	∆DFS>0	Se, Sp, Se+Sp, PPV, NPV	5	**YES**	**YES**
	Dou et al. (2018)	Cariogram (SC)	192 (10.7%)	18-29	2	∆DMFS>0	AUC, Se+Sp	5	**YES**	**YES**
Elderly	Petersson et al. (2003)	Cariogram (C)	148 (28%)	55, 65,75	5	∆DFS>0 ∆DS>0	None	4	**YES** **YES**	**YES** **YES**
	Hayes et al. (2017)	Cariogram (C)	280 (16.2%)	> 65	2	New root caries	AUC, Se, Sp, Se+Sp PPV,NPV	5	**YES**	**YES**

* Petersson (2010a) and Petersson (2010b) use the same sample but different versions of the Cariogram.
** These discrimination measures are not reported in Petersson (2010a) but in the companion article Petersson (2010b)

Table 2.

	Original	Platt	Isotonic
$Se._{05}$	86.90	83.07 (2.47)	85.12 (2.98)
$Sp._{05}$	87.61	89.60 (0.85)	88.46 (1.48)
$A^{0.5}$	12.65	12.77 (0.37)	12.76 (0.34)
AUC	0.93	0.93 (<0.001)	0.93 (0.001)
Calibration in-the-large	-0.3181	0.0011 (0.1162)	-0.0009 (0.1188)
Calibration slope	1.9733	0.9968 (0.0590)	0.9335 (0.0901)
ICI	0.1059	0.0210 (0.0040)	0.0157 (0.0062)
E50	0.0889	0.0151 (0.0038)	0.0113 (0.0073)
E90	0.1914	0.0494 (0.0178)	0.0396 (0.0192)
BS	0.1096	0.0952 (0.0005)	0.0944 (0.0008)